Data Quality

Table of Contents

INTRODUCTION

In every step of the data lifecycle, from collection to reporting, there is the potential for error. If your data has many errors, it will not give you an accurate picture of what's going on. Ensuring that you are collecting good quality, accurate data is vital to helping your government make informed decisions.

KEY POINTS IN THIS DOCUMENT

- Key criteria to determine the quality of your data sets are accuracy, consistency, completeness, uniqueness, timeliness, validity, and comparability.
- Think about the quality of data from a government-wide view, considering data across the organization, as well as the quality of individual data sets. For example, try to ensure that the same data point is not collected in multiple datasets.
- Proper documentation (metadata) is an important component of being able to understand the quality of your data and support its appropriate usage.
- Develop a quality assurance plan that outlines how data quality will be maintained throughout the data lifecycle.
- Data quality will be largely dictated by your data collection and input process. Key considerations for improving the quality of your data through modified data input processes include automation, standardization, integrated quality assurance, user experience, incentivization, and workflow.

DATASET QUALITY CRITERIA

How good is your data? How can you tell? Key criteria to determine the quality of an individual data set includes:

- **Accuracy**: Is information factual?
    - E.g., is an individual's birth date or address correct?
- **Consistency**: Are there rules and checks in place to ensure multiple users enter data in the same way? Are there processes that automate data entry to limit human choice and thus, error?
    - E.g., Are telephone numbers stored with or without dashes? Are dollar values stored in 100s or 1000s?
- **Completeness**: Is information collected for each individual and each variable?
    - E.g., Are there blank data points or individual records missing?
- **Uniqueness**: Are there any duplications in the data? Consider duplication in:
    - *Records:* is the same individual entered twice?
    - *Concepts:* is information conceptually duplicated within and across data sets (e.g., do you store both age and birth date?)
- **Timeliness**: Are data available frequently enough to be useful? Is data entered and processed within a reasonable timeframe?
- Relevance/**Validity**: Is the data relevant? Does it measure what it's supposed to measure?
    - E.g., Tax records are a more valid source of information on after-tax income than self-reported after-tax income.
- **Comparability:** If you are interested in looking at trends over time, is the data comparable year over year? If you are interested in comparing to other populations, is the data comparable to external data sets?

*Note: The criteria used to assess data quality are different, but related, to those used to assess indicator quality (see the document Indicator Identification and Framework).

DATA SYSTEM QUALITY

In addition to each of these quality measures, it is important to think of your data assets from a fuller, government-wide view (i.e., the data ecosystem). Even if each individual dataset you have is accurate, consistent, complete, unique, timely, valid, and comparable based on the above concepts, your government still may face data quality issues if:

- **Data is repeated across different data sets** (i.e., data is stored in more than one place). For example, you may store an individual's birth data in multiple data sets. If data is repeated, are there processes in place to update data across all data sets as needed, or to determine the authoritative version of the data?
    - For example, if you are storing a citizen's birthdate, and you need their age for a follow up analysis, the birthdate should be stored in a singular place, such as the citizenship database. Data from the citizenship database should then be extracted as appropriate in your follow up analysis. The citizenship database becomes the authoritative source of tombstone data (i.e., data that does not change over time) within your government and should be updated if anyone in the organization identifies an error.

    - When data is used by multiple users, it is important to clarify where the authoritative source for that data lies, rather than have each user store information separately.

- **Your government collects lots of data it doesn't use or need (**e.g., credit card information from former vendors or program participants.)
- **Information is not sensibly stored.** Datasets should be organized so that all the variables relate to each other and tangential or irrelevant data is stored separately. For example, if you are undertaking community planning exercises, community-level data describing the characteristics of a community (e.g., existing infrastructure), should be stored separately from citizen/individual level data describing the characteristics of an individual (e.g., income data). Although both may be relevant for community planning, organizing and storing them separately will improve the quality of your data.

DOCUMENTATION OF METADATA

Proper documentation is an important component of being able to understand the quality of your data and support its appropriate usage. Metadata is information about a dataset. It is important that metadata be consistently and comprehensively recorded so that:

1. There is consistency in how each data point is collected and entered.
2. Any potential user of the data will have the necessary context to understand and use it properly.

Metadata should include:

- **Who collected the dataset:** include contact information for questions.
- **A brief description of the dataset:** include format of the data, privacy concerns, quality concerns, completeness.
- **Where the dataset was collected:** the location the data was collected.
- **When the dataset was collected:** the timeframe of the data collection.
- **Why the dataset was collected:** what is the data to be used for.
- **How the data was collected:** refer to any documents outlining the data collection methodology.

Metadata for a dataset may be stored in a data dictionary. Data dictionaries contain detailed information about how each data point is collected and entered into a database.

QUALITY ASSURANCE

Quality assurance is the process of preventing errors in your data. It's a good idea to develop a **quality assurance plan** that outlines how data quality will be maintained throughout the data lifecycle, including[1]:

- Data quality objectives
- Requirements for:
    - Staff skills and training
    - Methods and equipment for data collection
    - Software and file types

[1] USGS. Manage Quality. https://www.usgs.gov/products/data-and-tools/data-management/manage-quality

    - Data standards
    - Regular data-quality assessments
- Data quality indicators
- Quality control processes for:
    - Reporting data errors
    - Checking for data errors and validating data
    - Data corrections
- Quality assurance documentation, including:
    - Data-quality assessment results
    - Staff training records
    - Data quality indicator outcomes

The process of entering data into a database is a key point in the data lifecycle where errors may be introduced. **Data entry quality assurance procedures** to help minimize errors and support a Quality Assurance plan include[2]:

- Develop a protocol for data entry.
- Ensure that each data entry field is labeled well.
- Create a data dictionary that gives clear instructions about what is to be entered in each data field.
- Define which fields are required vs optional. If a field is required, the data entry person will be forced to fill it in for the form to be labeled complete.
- Include fields in the database where the data entry person can indicate data quality. This could be a comment field where the data entry person can comment on unusual values or note known quality issues.

DATA ENTRY BEST PRACTICES

Data quality will be largely dictated by your data collection and input process. For instance, paper based and manual data entry are likely to lead to human error and require processes and standards to ensure consistency.

Some key considerations for improving the quality of your data through data input processes include:

- **Automation:** Data collection tools and database software can be used to reduce error, for example:
    - Having front line workers enter administrative data directly into the computer, rather than filling out a paper form limits the error that can be introduced during the process of transferring data from paper to computer.
    - Software may be used to auto-calculate things like age (using birthdate), date of record submission (using the date the record was created), or totals (by summing relevant fields).

[2] USGS. Quality by Design: Recommended practices. https://www.usgs.gov/products/data-and-tools/data-management/quality-design-recommended-practices?qt-science_support_page_related_con=0#qt-science_support_page_related_con

- **Standardization:** Data collection tools and database software can be used to limit field options, for example:
    - Use a data entry template to enter your data into the data management software. The template can restrict the type of data that can be entered into each field, limiting errors. For example, a field could be set to allow only letters, only numbers, or only dates.
    - Setting up data entry to select dates from a calendar rather than entering manually.
    - Drop down menus with fixed choices for particular fields.
- **Integrated Quality Assurance:** Use automatic quality controls in data management software to validate and audit data, for example:
    - Audit trails, i.e., a mechanism for tracking who generated which changes.
    - Sanity checks, i.e., checks within your data processing pipeline that ensure data is possible, consistent, and aligned with the codified expectations.
    - Validation tools, such as those that prevent letters from being entered in fields where only numbers should be stored.
    - Minimum and maximum allowable values for numeric fields so that impossible values cannot be entered. Ensure your minimum and maximum values are carefully identified and consider edge cases. You may also set a flag to indicate suspicious values that should be double checked by a data entry person.
- **User Experience:** Cumbersome, non-intuitive, or difficult to use data systems will have a large impact on the quality of data input. For example, a web form can be simpler for the user than entering data into excel directly.
- **Incentivization:** Think about how you can incentivize and enable your staff to take the time and attention to detail necessary to maintain data quality. See the Building Capacity - Change Management section of the toolkit.
- **Workflow:** Is data entry integrated into how staff already work? Data inputs that aim to be seamless with day-to-day tasks have a higher likelihood of improved data quality. For example, staff may enter information while seeing a client, rather than summarizing activities at the end of the day.

DATA QUALITY RESOURCES

The US Geological Survey website has good information and tools to manage data quality.

The US Geological Survey document "Metadata in plain language" includes a list of questions to consider when generating metadata.